Open

The Pharmacogenomics Journal (2010) 10, 375–384

www.nature.com/tpj

ORIGINAL ARTICLE

A *TOMM40* variable-length polymorphism predicts the age of late-onset Alzheimer's disease

AD Roses[1,2], MW Lutz[1,2], H Amrine-Madsen[3], AM Saunders[1,2], DG Crenshaw[1,2], SS Sundseth[1,2], MJ Huentelman[4], KA Welsh-Bohmer[1,5] and EM Reiman[4,6,7]

[1]*Department of Medicine, Duke University, Durham, NC, USA;* [2]*Deane Drug Discovery Institute, Durham, NC, USA;* [3]*GlaxoSmithKline, Research Triangle Park, NC, USA;* [4]*Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ, USA;* [5]*Alzheimer's Disease Clinical Center, Durham, NC, USA;* [6]*Banner Alzheimer's Institute, Phoenix, AZ, USA and* [7]*Department of Psychiatry, University of Arizona, Phoenix, AZ, USA*

Correspondence:
Dr AD Roses, Deane Drug Discovery Institute, School of Medicine and Fuqua School of Business, Duke University, R. David Thomas Executive Training Center, One Science Drive, Suite 342, Box 90344, Durham, NC 27708-0120, USA.
E-mail: allen.roses@duke.edu

Received 8 September 2009; revised 31 October 2009; accepted 2 November 2009; published online 22 December 2009

The ε4 allele of the apolipoprotein E (*APOE*) gene is currently the strongest and most highly replicated genetic factor for risk and age of onset of late-onset Alzheimer's disease (LOAD). Using phylogenetic analysis, we have identified a polymorphic poly-T variant, rs10524523, in the translocase of outer mitochondrial membrane 40 homolog (*TOMM40*) gene that provides greatly increased precision in the estimation of age of LOAD onset for *APOE* ε3 carriers. In two independent clinical cohorts, longer lengths of rs10524523 are associated with a higher risk for LOAD. For *APOE* ε3/4 patients who developed LOAD after 60 years of age, individuals with long poly-T repeats linked to *APOE* ε3 develop LOAD on an average of 7 years earlier than individuals with shorter poly-T repeats linked to *APOE* ε3 (70.5 ± 1.2 years versus 77.6 ± 2.1 years, $P = 0.02$, $n = 34$). Independent mutation events at rs10524523 that occurred during Caucasian evolution have given rise to multiple categories of poly-T length variants at this locus. On replication, these results will have clinical utility for predictive risk estimates for LOAD and for enabling clinical disease prevention studies. In addition, these results show the effective use of a phylogenetic approach for analysis of haplotypes of polymorphisms, including structural polymorphisms, which contribute to complex diseases.
The Pharmacogenomics Journal (2010) **10,** 375–384; doi:10.1038/tpj.2009.69; published online 22 December 2009

Keywords: AD genetics; phylogenetic analysis; *TOMM40*; *APOE*; poly-T variants

Introduction

The prevalence of Alzheimer's disease (AD) is predicted to quadruple worldwide by the year 2050 to >107 million cases, meaning that 1 in 85 persons will be living with the disease at that time. It has been estimated that delaying AD onset by 1 or 2 years could decrease the disease burden in 2050 by 9.5 million or 23 million cases, respectively.[1] Late-onset AD (LOAD), which develops after 60–65 years of age,[2,3] is the most common form of the disease, accounting for over 95% of cases.[3]

In all, 58–79% of the predisposition to LOAD is due to genetic factors.[4] Evidence gathered over the past 17 years clearly shows that apolipoprotein E (*APOE*) ε4 is the strongest and most highly replicated genetic risk factor for LOAD and is associated with lower age of clinical disease onset.[5,6] The early *APOE* discovery was based on modest human linkage data,[7] but has been followed by robust association of *APOE* genotypes with diagnosed LOAD and age of onset of onset distributions.[5] Several genome-wide association screens (GWAS)[6,8–15] and fine-mapping studies[16–18] have confirmed the association of a region of linkage disequilibrium (LD) that encompasses three genes, *APOE*, translocase of outer

mitochondrial membrane 40 homolog (*TOMM40*) and *APOC1*, with LOAD. However, despite the availability of whole genome single-nucleotide polymorphism (SNP) tools, no other common polymorphisms have shown similar robust and dramatic statistical associations with LOAD.

The rationale for the development of GWAS was to locate regions of the genome that may harbor specific disease-associated loci.[16,19] Recently, the emphasis has been on testing for association between SNPs that are represented on the available screening tools with a phenotype. This has introduced the statistical problem of correcting for the large numbers of SNPs assayed. A statistically significant association between an SNP and a phenotype tacitly grants priority to the SNP, when in fact the SNP is generally present on the commercial screening product only because it occurs at high frequency in some population and tags a region of LD. Even the most ardent enthusiasts for GWAS technology have been surprised and disappointed by the lack of robust disease-specific results, leading to calls for analyses of combined series of tens of thousands of patients and controls.[20]

We approached the LD region that includes *APOE* and *TOMM40* by deep sequencing to catalog all the polymorphisms, including structural polymorphisms in addition to SNPs, and then applying phylogenetics to define the evolutionary relatedness of the polymorphisms. This technique is used extensively for evolutionary analyses, from the evolution of species to the changes occurring in influenza virus each year. Phylogenetics has been used less frequently for human disease genetics, but is ideally suited for analysis of regions of the genome where there is high sequence diversity and low levels of recombination. Phylogenetic analysis is fundamentally different from GWAS in that it is not searching for disease-associated chunks of DNA that represent LD regions, but rather it identifies collections of related haplotypes with common ancestral history, that is clades, that may be enriched for disease-causing variants. Preliminary genome-wide screens are, therefore, valuable for flagging linkage regions of potential interest for a particular phenotype. Using a phylogenetic analysis of a previously flagged genomic region,[16] we have discovered a polymorphic poly-T variant in *TOMM40* that is linked to *APOE*. The length of this polymorphic poly-T contributes to the age of onset distributions formerly attributed to *APOE* genotypes alone by making *APOE* ε3-containing strands more informative.

Mitochondrial dysfunction is an early defect in LOAD pathogenesis[21–27] and is linked to neuronal cell death.[28] One candidate gene for mitochondrial dysfunction in LOAD is *TOMM40*. This gene encodes Tom40, the translocase of the outer mitochondrial membrane pore subunit, through which cytoplasmic peptides and proteins pass during mitochondrial biogenesis.[29] Amyloid precursor protein has been shown to accumulate in the mitochondrial import pores, which results in mitochondrial dysfunction in LOAD.[30,31] In addition, mitochondrial dysfunction and neurotoxic effects of naturally occurring, neuron-specific apoE4 1-272 N-terminal peptide fragments interacting at the outer mitochondrial membrane have also been described.[28] The 3′ and 5′ ends of the *TOMM40* and *APOE* genes, respectively, are separated by only ~2 kb on chromosome 19. The *TOMM40* and *APOE* genes are in high LD,[17,18] which may obscure disease risk associated with other *APOE* ε4-independent variants in the region.

Phylogenetic analysis has been used previously to identify genomic relationships between low-frequency genetic variants and to cluster evolutionarily related haplotypes.[32] In this study this methodology is used to explore the *TOMM40-APOE* LD block for the existence of novel risk determinants for LOAD.

Materials and methods

Subjects

The two cohorts analyzed in this study were from the Arizona Alzheimer's Disease Research Center, Phoenix, Arizona, and the Duke Bryan Alzheimer's Disease Research Center, Durham, North Carolina. Details of the Exploratory Study cohort are given in Li *et al*.[6] All subjects were of European descent. The Arizona and Duke studies were approved by the institutional review boards and appropriate informed consent was obtained from all participants. Age and gender data for the cases and controls in each cohort are shown in Table 1. For the Duke cohort, the age of disease onset was determined retrospectively and disease diagnosis was confirmed by autopsy for subjects who have died.

DNA samples

For the Exploratory study (ES) and Arizona study (AS) cohorts, DNA was extracted from blood. For the Duke study (DS) cohort, DNA was extracted from the blood (for 22 subjects) or brain (for 12 subjects). There was no systematic bias for the tissue of origin of the DNA in the final analysis of the DS cohort, that is, long and short rs10524523 alleles were found in DNA from both tissues. Samples were plated on 96-well plates for long-range PCR and DNA sequencing at Polymorphic DNA Technologies (Alameda, CA, USA).

Long-range PCR

Long-range PCR was performed using Takara LA Taq Polymerase (Takara Mirus Bio, Inc., Madison, WI, USA). The reaction mix and PCR conditions were the same as those recommended by the manufacturer. PCR was conducted in

Table 1 Cohort compositions

Series	n		*Mean age (s.d.)*		*Females (%)*	
	Cases	*Controls*	*Cases*	*Controls*	*Cases*	*Controls*
ES	83	67	72.0 (0.9)	74.6 (3.3)	60.2	59.7
AS	74	31	81.7 (8.0)	77.0 (8.9)	56.3	46.7
DS	34	33	69.3 (8.3)	71.9 (7.5)	70.0	66.7

Abbreviations: AS, Arizona study; DS, Duke study; ES, Exploratory study.
The number of cases and controls, mean age and percentage that are female are shown for each series. Mean age is given as age-at-diagnosis of Alzheimer's disease for cases and age-at-examination for controls. The s.d. from the mean is given in parenthesis.

a 50 μl volume with 2.5 U of LA Taq and 200–400 ng human genomic DNA. Thermocycling was carried out with the following conditions: 94 °C, 1 min for 1 cycle; 94 °C, 30 s; 57 °C, 30 s; 68 °C, 9 min for 14 cycles; 94 °C, 30 s; 57 °C, 30 s; 68 °C, 9 min + 15 s per cycle for 16 cycles; 72 °C, 10 min for 1 cycle. Primers for long-range PCR are shown in Supplementary Table S1.

Large-fragment cloning
PCR products were run on a 0.8% agarose gel, visualized by crystal violet dye, compared with size standards, cut out of the gel and extracted with purification materials included with the TOPO XL PCR Cloning kit (Invitrogen, Carlsbad, CA, USA). Long-range PCR products were cloned into a TOPO XL PCR cloning vector. This system uses a TA cloning vector and is recommended for inserts of up to 10 kb. As per the manufacturer's instructions, electro-competent cells (from the same kit) were transformed by the vector, plated in the presence of antibiotic and incubated. Altogether, 10 clones from each plate were picked and cultured in a 96-well format.

Template preparation
Diluted cultures were transferred to a denaturing buffer that was part of the TempliPhi DNA Sequencing Template Amplification kit (GE Healthcare/Amersham Biosciences, Piscataway, NJ, USA). This buffer causes the release of plasmid DNA but not bacterial DNA. Cultures were heated, cooled, spun, and transferred to fresh plates containing the TempliPhi enzyme and other components. This mixture was incubated at 30 °C for 18 h to promote amplification of the plasmid templates. These products were then spun and heated to 65 °C to destroy the enzyme.

DNA sequencing
Plasmid templates were used in DNA sequencing reactions using the Big Dye, version 3.1 sequencing kit (Applied Biosystems, Foster City, CA, USA). For each reaction, an appropriate sequencing primer (Supplementary Table S1) was used that was designed to anneal to a unique location of the template. Cycle sequencing was carried out with an annealing temperature of 50 °C, an elongation temperature of 60 °C and a denaturation temperature of 96 °C, for a total of 30 cycles. Sequencing reaction products were run on an ABI 3730XL DNA sequencer with a 50-cm capillary array using standard run mode.

Sequencing data analysis
A proprietary sequencing analysis program called 'Agent' (developed by Celera, Alameda, CA, USA) was used to align sequencing reads to the appropriate reference sequence, and produce 'contigs' associated with each clone. The system provides estimated quality scores for all bases for which there is any variation for any of the samples. The sequencing report for each sample was analyzed for the presence of SNPs that were correlated in one haplotype pattern for one subset of clones and in a different haplotype pattern for the remaining clones. A reference file for the region of interest was prepared by listing the known variations for that region publicly available from NCBI dbSNP. A genotype file for the region of interest was created by searching each subject's haplotype report for all variations between the known reference sequence and the consensus haplotype sequences.

Estimate of length-read error
The magnitude of the length-reading error for the poly-T variants (for example, rs10524523) was estimated by examining the observed lengths from the 10 clones that were prepared for samples that had a single haplotype. For a typical sample with short poly-T length of 16, the s.d. for the 10 clones was 0.97. For a typical sample with longer poly-T length, for example, 27, the s.d. was 1.58.

Phylogenetic analysis
Phylogenetic analysis was conducted according to the steps outlined in Supplementary Figure S1. A multiple sequence alignment of the sequences was performed using the ClustalW2 (version 2.0.10, European Bioinformatics Institute (EBI), Hinxton, UK; http://www.ebi.ac.uk/Tools/clustalw2/index.html) program using default parameters. Manual adjustment of the alignments was completed using Genedoc (version 2.7.000, National Resource for Biomedical Supercomputing (NRBSC), Pittsburgh, PA, USA; http://www.nrbsc.org/gfx/genedoc/index.html). Phylogenetic trees were constructed using Bayesian, maximum likelihood and distance-based reconstructions. The phylogenetic tree construction software used was Paup* (version 4.0b10, Sinauer Associates, Sunderland, MA, USA; http://paup.csit.fsu.edu), ClustalW2 (neighbor-joining methods, version 2.0.10, European Bioinformatics Institute (EBI), Hinxton, UK; http://www.ebi.ac.uk/Tools/clustalw2/index.html) and MrBayes (version 3.1.2, Florida State University, Tallahassee, FL; http://mrbayes.csit.fsu.edu/index.php).

Tree-bisection and reconnection branch swapping were used in all methods. The best fitting model of sequence evolution was estimated using the Modeltest program (version 3.7, University of Vigo, Spain; http://darwin.uvigo.es/software/modeltest.html), which provided estimates for the following key determinants: rate matrix, shape of the gamma distribution and proportion of invariant sites. Bootstrap analysis was performed using 1000 replicates to determine statistical support for specific tree morphology.

Haplotype networks were also constructed from the sequence data using the program TCS (version 1.21[33], University of Vigo, Spain; http://darwin.uvigo.es/software/tcs.html) to compare the phylogenetic trees to cladograms estimated using statistical parsimony. The phylogenetic trees and haplotype networks were constructed twice, with gaps treated as missing data for the first instance and as a fifth character for the second instance. Nucleotide diversity in the region of interest was calculated using DnaSP (version 5.00.02[34], University of Barcelona, Spain; http://www.ub.edu/dnasp).

After construction of the phylogenetic trees, the haplotype network and completion of the analysis of nucleotide diversity in the region of interest, the results from the

different methods were compared and reconciled to a consensus tree. Groups of sequences sharing a recent disease mutation were presumed to segregate more closely on the phylogenetic tree; however, sporadic cases due to phenocopies, dominance and epistasis can introduce noise into the phenotype–haplotype relationship.[35] This phylogenetic analysis focused on a high-level aggregation of clades to minimize these effects. The clades determined at the first split in the phylogenetic tree were used to test the hypothesis that *TOMM40* subject haplotypes from clade 'B' were associated with the onset of AD at a later age than subject haplotypes from clade 'A' (each subject contributed two haplotypes to the AD age of onset association signal). The number of tests of association that are performed using this approach was orders of magnitude less than that in typical GWAS, as the phylogenetic analysis identified categories of evolutionarily related subject haplotypes. If the tests of association confirmed that the different clades classified the subject-haplotype data by age of onset, further statistical analysis was carried out to identify the variants that separated the sequences into each clade. Effectively, this analysis assessed the significance of each variant as a factor that influences age of onset using a series of one degree of freedom tests guided by the tree structure. The phylogenetic analyses were conducted using SNP and insertion/deletion polymorphisms. The statistical tests of association were adjusted with a Bonferroni correction for the number of polymorphic sites included in the analysis.

Statistical analyses

Haplotype reports from the Polymorphic analysis software and reports from DnaSP software (version 5.00.02[34]) were used for subsequent statistical analyses. We analyzed individual *TOMM40* SNP variants, *TOMM40* haplotypes and length of poly-T repeats for association with LOAD risk for the AS cohort and LOAD age of onset for the DS cohort. Differences in the proportions of specific *TOMM40* alleles associated with each *APOE* allele or *APOE* genotype were compared using Fisher's exact test (two-tailed). Starting with 30 parsimony-informative sites and $\alpha = 0.05$, a Bonferroni correction for the significance of a specific allelic association would require a *P*-value of 0.001. Odds ratios (ORs) were calculated as the (number of minor alleles in cases/number of minor alleles in controls)/(number of major alleles in cases/number of major alleles in controls) and reported with 95% confidence interval. Means for defined LOAD age of onset groups were compared by *t*-tests (two-tailed). A standard F-test on group variances was performed to determine whether the *t*-test was calculated assuming equal or unequal variances. Statistical analysis was completed using JMP software (version 8, SAS Institute, Cary, NC, USA).

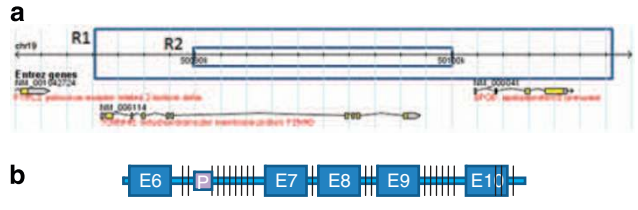

Figure 1 Schematic overview of the *TOMM40-APOE* locus. (**a**) Genomic locations of the *TOMM40* and *APOE* genes on chromosome 19 between 50 084 480 and 50 107 480 bp. The regions subjected to primary sequencing and phylogenetic analysis for the exploratory (R1) (23 kb) and confirmatory (R2) (10 kb) studies are highlighted on the genomic map (NCBI Build 36.3). (**b**) Distribution of SNP and insertion/deletion polymorphisms are shown on the gene structure of *TOMM40* covered in the region that was subjected to primary sequencing. The region covers exons 6–10 and all associated intronic regions. The variable poly-T repeat (rs10524523) that is significantly associated with LOAD age of onset is depicted with the square labeled 'P'.

Results

Molecular evolutionary analysis of the TOMM40-APOE *locus*

In an ES, 23 kb of DNA containing the *TOMM40* and *APOE* genes (R1 in Figure 1a) was amplified and sequenced for 83 LOAD cases and 67 age-matched controls, and included subjects with *APOE* ε3/3, ε3/4 and ε4/4 genotypes (no ε2 alleles) (details of the ES cohort are given in Li *et al*.[6]). The 23-kb region is part of an extended region of LD containing *APOE*[17,18] and was selected because it fully contained both the *TOMM40* and *APOE* genes plus almost 3 kb of flanking sequencing on either side and because of earlier reports that the *TOMM40* gene may be involved in LOAD pathogenesis.[6,16] To accomplish sequencing, the 23-kb genomic region was divided into three ~10-kb overlapping segments (Supplementary Figure S2). Molecular evolutionary analyses of the three 10-kb regions included phylogenetic reconstruction, statistical parsimony, haplotype networks and polynucleotide repeat analysis. Of the three segments, only one (R2, Figure 1a) supported a phylogenetic tree structure that had high bootstrap support. The R2 segment encodes *TOMM40* exons 6–10 (Figure 1b). The phylogenetic topology of the R2 segment had three notable characteristics. First, the phylogenetic tree was divided into two major clades (termed A and B) with strong bootstrap support (98%, 1000 replicates). This is an unusually large distinction for the human intraspecific data. Second, phase resolution of haplotypes of *TOMM40* polymorphisms in R2 with the linked *APOE* ε4 or ε3 allele revealed a highly significant difference in the distribution of the *APOE* alleles and genotypes and *TOMM40* haplotypes between clades A and B on the tree. Both clades contained subjects with the ε3/3 genotype, but 98% of all clade B *TOMM40* haplotypes occurred in *cis* to the *APOE* ε3 allele ($P = 2.0 \times 10^{-24}$, $n = 300$, Fisher's exact test, two-tailed) (Figure 2). Third, the *TOMM40-APOE* genomic region contained a large number of polynucleotide repeats; however, the repeats that were polymorphic in length were concentrated in intron 6 of *TOMM40*, which is included in the R2 region. Taken together, these data suggested that this 10-Kb R2 region of *TOMM40* was functionally significant and that these *TOMM40-APOE* haplotypes could account for the robust genetic association of this high LD region.

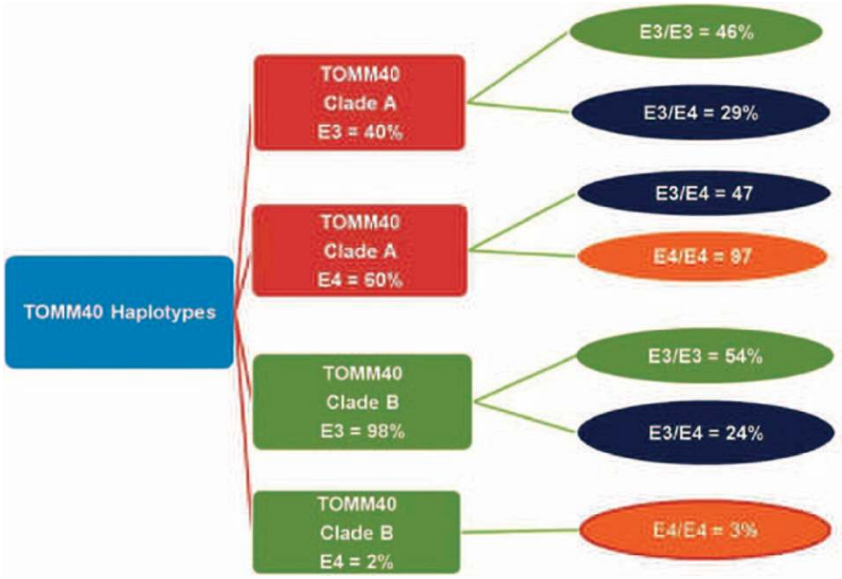

Figure 2 Schematic of *APOE* allele and genotype frequencies associated with the major clades for the phylogenetic tree constructed for the *TOMM40* R2 region (Figure 1) for the ES cohort. The percentage of each *APOE* allele that is present in clades A and B is indicated (column 2). The *APOE* genotypes that are represented in each clade are also given (column 3). $n = 300$ subject haplotypes.

The phylogenetic structure of the 10-kb region (R2), the *APOE* ε3-specific inheritance of *TOMM40* haplotypes from clade B, the variable-length polynucleotide repeats and the identity of the clade-specific polymorphisms that were observed in the ES were all confirmed in a second independent LOAD case–control cohort. The AS cohort was comprised of AD cases ($n = 74$) and controls ($n = 31$) with *APOE* genotypes ε2/3, ε2/4, ε3/3, ε3/4 and ε4/4, and the subjects were ascertained at the Arizona Alzheimer's Disease Core Center. The association between the two clades, disease risk and age of disease onset was explored in the AS cohort that had a broader spectrum of ages of onset than the ES. A third cohort, the DS, was assembled at the Duke Bryan Alzheimer's Disease Research Center and comprised 34 clinically well-characterized *APOE* ε3/4 LOAD patients, many of whom had AD confirmed by autopsy on death, and 33 age-matched controls. Table 1 summarizes the characteristics of the three cohorts.

Reconstruction of the evolutionary history of the AS cohort revealed a highly similar phylogenetic tree to that seen in the ES, with strong bootstrap support (97%, 1000 replicates) for the separation of clades A and B. *APOE* ε4/4 subjects occurred only in clade A, whereas the remaining *APOE* genotypes were distributed between clades A and B (Supplementary Figure S3). Examination of the distribution of the few *APOE* ε2/4 subjects on the phylogenetic tree suggested that *APOE* ε2-*TOMM40* haplotypes share a similar evolutionary history with *APOE* ε3-*TOMM40* haplotypes. To verify the phylogenetic structure using a separate method, and to ensure that recombination within the genetic interval did not confound the phylogenetic tree structure, haplotype networks were also constructed using statistical parsimony (TCS version 1.21[33]). The major subject-haplotype clusters derived from the two methods (maximum parsimony and TCS) were congruent.

Association of the rs10524523 poly-T polymorphism with LOAD risk and age of onset

Mapping the polymorphisms that distinguished the two major clades of the phylogenetic tree derived for the AS cohort showed that a poly-T variant, rs10524523, located in intron 6 of *TOMM40* was a key variant that separated the two clades and, therefore, the two groups of *APOE* ε3 haplotypes. For *APOE* ε4/4 subjects, the variant was relatively long with a narrow, unimodal distribution of lengths (21–30 T residues, mean = 26.78, s.d. = 2.60, $n = 32$), whereas for *APOE* ε3/3 subjects, a bimodal distribution of lengths was evident with peaks at 15.17 (s.d. = 0.85, $n = 36$) and 33.15 (s.d. = 2.09, $n = 55$) T residues (Figure 3a). Two *APOE* ε4/4 AD patients, each having a short poly-T allele (length 15) that mapped to clade B on the phylogenetic tree, were identified. These two patients had a later age of AD onset (78 years) than would be expected, on average, for individuals possessing two *APOE* ε4 alleles. Histograms of the rs10524523 length distributions and allele frequencies for control subjects (that is, not representative of the general population) from the ES cohort are also provided in Supplementary Figure S4.

There was a significant association between length category of the rs10524523 poly-T polymorphism and age of LOAD onset for the DS cohort of *APOE* ε3/4 subjects, for whom disease-onset data were available. Longer poly-T alleles ($\geqslant 27$ T residues) were significantly associated with the onset of disease at a much younger age than shorter poly-T alleles (70.5 years ± 1.2 versus 77.6 years ± 2.1, $P = 0.02$, $n = 34$). This polymorphism, therefore,

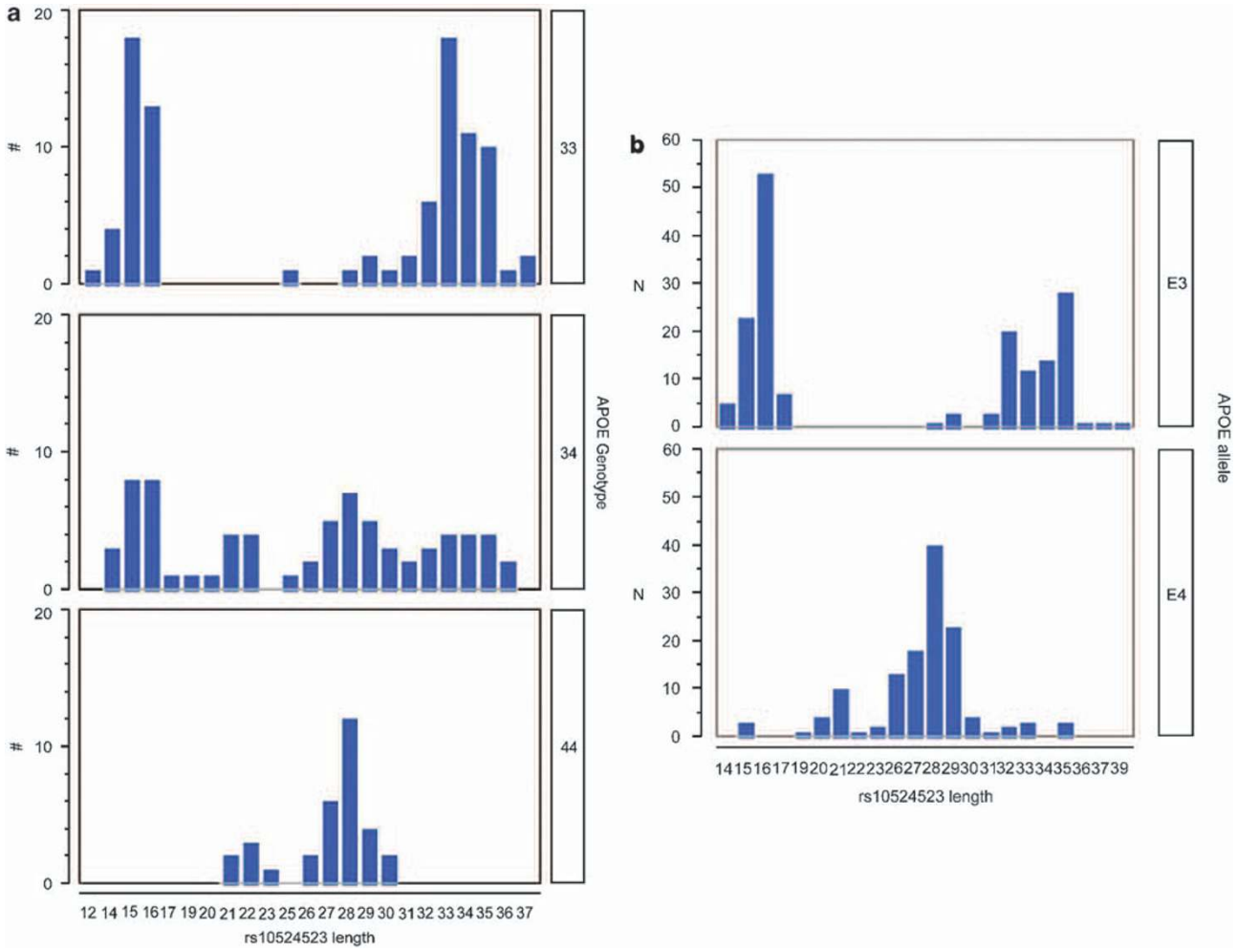

Figure 3 Distributions of the rs10524523 variable-length polymorphism for the AS ($n=210$ subject haplotypes) and ES ($n=300$ subject haplotypes) cohorts. (**a**) Histograms of rs10524523 length stratified by *APOE* genotype for the AS cohort. (**b**) Histograms of rs10524523 poly-T lengths and allele frequencies linked to specific *APOE* alleles for the ES cohort.

significantly impacted age of disease onset for individuals who carried a single *APOE* ε3 allele independent of carriage of a single ε4 allele. Three other poly-T length polymorphisms located in intron 6 (rs34896370, rs56290633 and rs10602329) also distinguished clades A and B, but these polymorphisms were not associated with a statistically significant difference in age of disease onset.

Longer poly-T lengths (T $\geqslant$27) segregated almost exclusively into clade A, the higher disease risk clade, in the AS cohort ($P=7.6\times10^{-46}$, $n=210$, Fisher's exact test, two-tailed). The distributions of poly-T lengths linked to specific *APOE* alleles in subjects from the ES cohort are shown in Figure 3b, confirming the *APOE* allele-specific distribution of poly-T lengths for the ES and AS cohorts. Case–control ratios for rs10524523 poly-T lengths for the AS cohort are provided in Supplementary Table S2.

TOMM40 haplotypes and variants associated with LOAD

AD cases more frequently possessed clade A haplotypes than clade B haplotypes for all study cohorts (for example, for the AS cohort, OR = 1.44, 95% confidence interval = 0.76–2.70). *APOE* ε3/4 heterozygotes in the AS cohort ($n=36$) were analyzed to estimate disease risk associated with clade A haplotypes while controlling for the effect of *APOE* ε4. There was a trend to higher incidence of LOAD for the subset that was homozygous for *TOMM40* clade A haplotypes relative to the subset that was heterozygous for clade A and clade B haplotypes (OR = 1.36, 95% confidence interval = 0.40–4.61), and thus it is possible that other variants in *TOMM40* that define clade A, in addition to rs10524523, confer *APOE* ε4-independent risk of LOAD.

Analysis of the AS cohort sequence data identified 39 polymorphic sites, including the poly-T sites, in the *TOMM40* 10-kb R2 region. In total, 10 SNPs occurred exclusively in the context of *APOE* ε3 ($P=6.07\times10^{-50}$, $n=210$, Fisher's exact test, two-tailed) and were never observed in *APOE* ε4/4 homozygous subjects ($n=16$). Figure 4 shows 16 SNPs that distinguish *TOMM40* clades A and B for the *APOE* ε3/3 subjects from the AS cohort. These polymorphisms were tested individually and

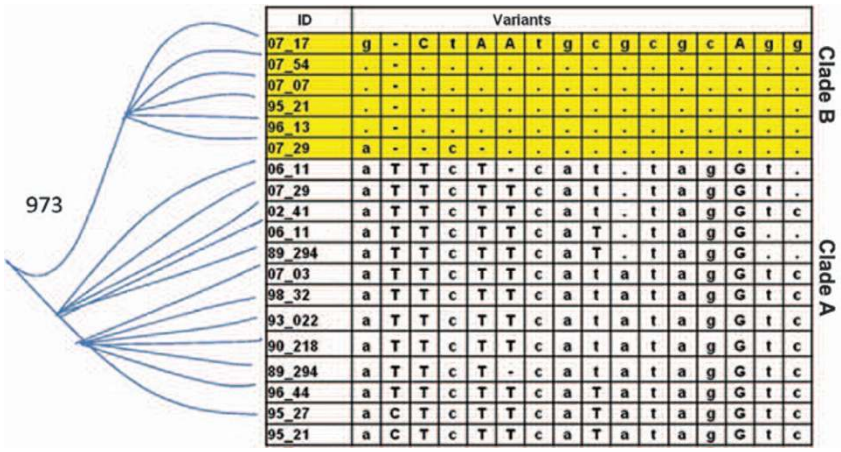

Figure 4 Phylogenetic tree showing separation of SNP variants that distinguish clades A and B. Separation of the two main branches has strong bootstrap support (973 of 1000). These SNPs distinguish the two clades for both the ES and AS cohorts.

as haplotypes for association with LOAD risk (Supplementary Table S3). The ORs for disease risk for each clade B allele, in all cases the minor allele, suggested that the clade B alleles are protective of AD risk in the AS cohort, although in each case the association narrowly missed significance. However, the minor alleles of four of the SNPs (rs8106922, rs1160985, rs760136 and rs741780) that distinguish *TOMM40* clade B were assayed previously in four published LOAD case–control GWAS and were found to be significantly protective of disease risk (OR <1 in each case), which is consistent with the trend observed in our smaller study.[8,11,15,17]

Discussion

Our discovery of a variable-length sequence repeat polymorphism associated with LOAD age of onset adds to the list of previous examples in which the sequence repeat length affects disease risk or disease penetrance. For example, unstable trinucleotide repeats are known to cause or contribute to risk of at least 14 neurological diseases.[36] For Huntington's disease, chromosomes containing short (<36) polyglutamine (poly-CAG) repeats are benign; however, an increase of only 2–3 repeats is associated with increased disease risk.[37] Repeat length is also inversely correlated with the age of onset of Huntington's disease.[38] Common poly-T polymorphisms and TG-repeats in the *CFTR* gene have been shown to have a role in the development of cystic fibrosis-related diseases.[39] The IVS8 poly-T alleles located in the splicing acceptor site of intron 8 of *CFTR*, for example, variously affect skipping of exon 9 and the production of nonfunctional protein.[40] Whether different poly-T lengths at *TOMM40* rs10524523 also result in exon skipping is unknown. Alternatively, it is possible that the rs10524523 polymorphism, alone or in conjunction with other SNPs in *TOMM40*, acts at a distance to affect transcription of *APOE*. It has previously been shown that polymorphisms in *TOMM40* affect levels of apoE protein in the cerebrospinal fluid of non-demented individuals and in the hippocampus of AD patients.[41,42] Finally, as there is evidence of interaction between apoE protein and the mitochondria,[28] it could be envisioned that the *TOMM40* polymorphisms affect the production of specific Tom40 isoforms that interact in a complex way with apoE isoforms encoded by *APOE* ε2, ε3 or ε4.

It is highly probable that African, Asian, Caucasian and other ethnic groups have very different phylogenetic patterns in the *APOE-TOMM40* region. This may affect the clinical usefulness, for non-Caucasians, of the data presented here and this could be especially problematic in the pharmacogenetic interpretation of global clinical trials. This factor must be considered when large Phase III trials do not confirm the efficacy found in original Phase II experiments based solely on Caucasians. As an example, it has been established that treatment of non-small cell lung carcinoma with gefitinib is particularly efficacious for Asian patients relative to Caucasian patients and this is related to mutations of the drug target that occur frequently in Asians.[43,44] In addition, there are descriptions in the literature of ethnic-specific adverse events.[45,46] The linkage of potentially contributing loci, which may vary according to the evolutionary history of different ethnicities, should be considered carefully when designing large, global clinical trials in which pharmacogenetics will be applied. There are certainly data sets where this could be tested, as it may be a contributing factor to the recent drug development failures in which clinical trials had different ethnic mixes.[47–49]

We conclude that longer poly-T tracts at rs10524523 are significantly correlated with earlier age of onset of LOAD. The length of this variant on *APOE* ε4 chromosomes is relatively homogeneous and relatively long on *APOE* ε4 chromosomes, whereas there are two distinct groups of poly-T lengths linked to *APOE* ε3. *APOE* ε2 chromosomes also seem to carry variable-length poly-T repeats similar to ε3 chromosomes, but further investigation is needed to verify this preliminary finding and to determine whether the poly-T repeat affects the very late age of disease onset for carriers of *APOE* ε2. Of all the variants that distinguished the

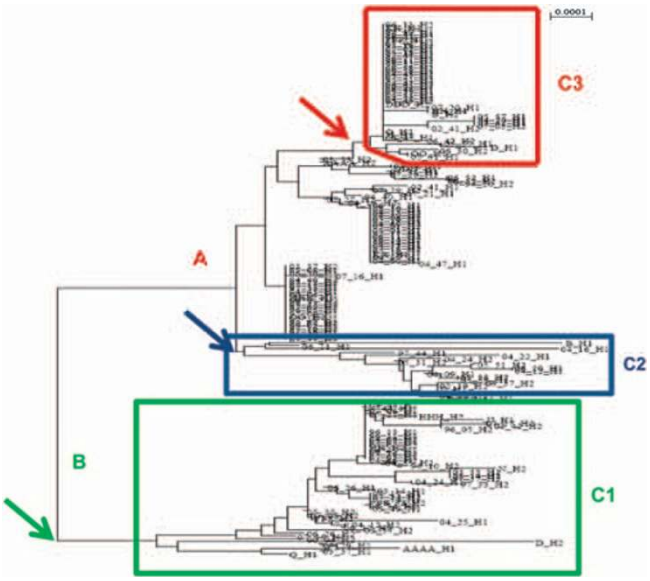

Figure 5 Annotated phylogenetic tree shows that there are three groups of clades on the tree that have consistent and characteristic patterns of rs10524523 lengths (C1-C3). For the major clade B, which is associated with lower risk of LOAD and contains haplotypes linked to the *APOE* ε3 allele, rs10524523 lengths are characterized by a uniform short length (C1: mean = 15.9, s.e.m. = 0.4, $n = 63$). There are two groups that branch from major clade A, one group is characterized by short lengths of rs10524523 (C2: mean length = 21.4, s.e.m. = 0.2, $n = 16$) and the second by long rs10524523 lengths (C3: mean length = 33.5, s.e.m. = 0.3, $n = 30$). Haplotypes in group C3 are linked uniquely to the *APOE* ε3 allele. The remaining subject haplotypes in clade A have a mean poly-T rs10524523 length of 30.2 (s.e.m. = 0.3, $n = 101$) and are associated with both *APOE* ε3 and ε4 alleles. There is a strong bootstrap support (973/1000) for the first major branch and moderate bootstrap support for the branches within clade A (247 of 1000 for C3 and 776 of 1000 for C2). Arrows indicate the branching point for each of the groups.

two major clades of the phylogenetic tree, rs10524523 best describes age of LOAD onset. However, at this time, we cannot exclude the possibility that other variants that occur in haplotypes with rs10524523, or in the *APOE* LD region which we did not sequence are actually causative. The distribution of rs10524523 of length alleles into major clades of the phylogenetic tree suggests that this polymorphism is inherited faithfully in haplotypes with specific alleles of *APOE* (Figure 5) and do not represent dynamic mutations as observed in other neurological diseases. Although it is possible that there are other variants that influence the age of onset of LOAD for individuals who are not homozygous for *APOE* ε4, the length of the poly-T polymorphism in *TOMM40* intron 6 seems to be the most powerful genetic predictor in this linkage region, and should be validated in prospectively collected series of normal subjects being followed for the development of LOAD. These data also suggest that *APOE* genotype-stratified age of onset curves[5,6] are, in reality, families of curves with each curve reflecting a specific interaction of linked polymorphisms in *APOE* and *TOMM40*. After validation, these data will add resolution to the prediction of age of LOAD onset, within an 8-year window, for Caucasian individuals >60 years of age. A prospective, population-based study to validate the association of *APOE* genotypes and *TOMM40* haplotypes or rs10524523 with age of disease onset, and to determine the utility of these alleles for prediction of age of onset, is currently being planned. This study will be a prospective, 5-year population-based study conducted in several ethnic groups, and will be combined with a prevention or delay of disease onset drug trial for individuals whose genetics and age would predict that they are at high risk of developing LOAD within 5–7 years. Population-based studies should also be conducted in several ethnic groups for whom the phylogenetic structure of the *TOMM40-APOE* region is known to further our understanding of the role of the poly-T repeat and other nearby polymorphisms in LOAD.

Accession codes
GenBank: *TOMM40* translocase of outer mitochondrial membrane 40 homolog, 10452; *APOE* apolipoprotein E, 348.

Conflict of interest

Dr Roses is the President of Zinfandel Pharmaceuticals, a single-owner corporation in the State of North Carolina that is planning the diagnostic validation study for the poly-T repeat variant as an age-dependent risk of AD in geography-based populations. This study will be coupled with a drug-prevention/delay of onset clinical trial. Pharmaceutical partners with suitable molecules are being evaluated over the next 6–12 months, while the experimental design of the study is reviewed by the FDA VxDS process, and while the recruitment and organization of study sites is finalized (http://opalstudy.org/index.html). Simultaneously, a consultation group of ethical, legal and social experts was convened in October 2008, and has been examining these questions before the diagnostic covered by submitted patents will be licensed for commercial uses. This paper provides the data and the first opportunity for other research laboratories to test and validate the data. This activity is independent of the Duke University, but the intellectual property is intended to be treated as Deane Drug Discovery Institute property once there is an established commercial value. All the remaining coauthors declare no conflict of interest.

Acknowledgments

We thank Dr Ornit Chiba-Falek, Kathleen M Hayden, Sandy Stinnett, Elizabeth Harris, April N Allen and Jason J Corneveaux for their scientific and technical contributions to this work. The authors would like to recognize the contribution of DNA samples from the Netherlands Brain Bank (under the direction of Dr Rivka Ravid) and Banner Sun Health Research Institute (under the direction of Dr Thomas Beach). Work at the Arizona ADCC was supported in part by the National Institute on Aging Grants P30 AG19610 and R01 AG031581 to (to EMR), a National Institute of Neurological Disorders and Stroke Grant R01 NS059873 (to MJH), a Science Foundation Arizona grant (to MJH), the Arizona Alzheimer's

Consortium and the State of Arizona. The work at the Bryan ADRC was supported in part by the NIA Grant AG028377. Dr. Roses and Dr. Lutz are supported in part by a grant from the NIA (1RC1 AG03563501). This research was also supported by a gift from an anonymous donor and by the Deane Drug Discovery Institute. We thank the donors, families and caregivers for their participation in this effort.

References

1 Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement* 2007; **3**: 186–191.
2 Pericak-Vance MA, Grubber J, Bailey LR, Hedges D, West S, Santoro L *et al.* Identification of novel genes in late-onset Alzheimer's disease. *Exp Gerontol* 2000; **35**: 1343–1352.
3 Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimers Dement* 2008; **4**: 110–133.
4 Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S *et al.* Role of genes and environments for explaining Alzheimer's disease. *Arch Gen Psychiatry* 2006; **63**: 168–174.
5 Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 1993; **261**: 921–923.
6 Li H, Wetten S, Li St L, Jean PL, Upmanyu R, Surh L *et al.* Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer's disease. *Arch Neurol* 2008; **65**: 45–53.
7 Pericak-Vance MA, Bebout JL, Gaskell Jr PC, Yamaoka LH, Hung WY, Alberts MJ *et al.* Linkage studies in familial Alzheimer's disease: evidence for chromosome 19 linkage. *Am J Hum Genet* 1991; **48**: 1034–1050.
8 Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L *et al.* A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Med Genomics* 2008; **1**: 44.
9 Beecham GW, Martin ER, Li Y-J, Slifer MA, Gilbert JR, Haines JL *et al.* Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer's disease. *Am J Hum Genet* 2009; **84**: 35–43.
10 Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF *et al.* Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. *Am J Hum Genet* 2008; **83**: 623–632.
11 Carrasquillo MM, Zou F, Pankratz VS, Wilcox SL, Ma L, Walker LP *et al.* Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nat Genet* 2009; **41**: 192–198.
12 Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Lince DH *et al.* A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry* 2007; **68**: 613–618.
13 Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A *et al.* Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. *Hum Mol Genet* 2007; **16**: 865–873.
14 Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* 2009; **41**: 1094–1099.
15 Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, Fallon JH *et al.* Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One* 2009; **4**: e6501.
16 Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J *et al.* SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer's disease. *Am J Hum Genet* 2000; **67**: 383–394.
17 Takei N, Miyashita A, Tsukie T, Arai H, Asada T, Imagawa M *et al.* Genetic association study on in and around the APOE in late-onset Alzheimer's disease in Japanese. *Genomics* 2009; **93**: 441–448.
18 Yu C-E, Seltman H, Peskind ER, Galloway N, Zhou PX, Rosenthal E *et al.* Comprehensive analysis of APOE and selected proximate markers for late-onset Alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. *Genomics* 2007; **89**: 655–665.
19 Lai E, Riley J, Purvis I, Roses A. A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* 1998; **54**: 31–38.
20 Goldstein DB. Common genetic variation and human traits. *N Engl J Med* 2009; **360**: 1696–1698.
21 Crouch PJ, Cimdins K, Duce JA, Bush AI, Trounce IA. Mitochondria in aging and Alzheimer's disease. *Rejuvenation Res* 2007; **10**: 349–358.
22 Atamna H, Frey WH. Mechanisms of mitochondrial dysfunction and energy deficiency in Alzheimer's disease. *Mitochondrion* 2007; **7**: 297–310.
23 Wang X, Su B, Zheng L, Perry G, Smith MA, Zhu X. The role of abnormal mitochondrial dynamics in the pathogenesis of Alzheimer's disease. *J Neurochem* 2009; **109**: 153–159.
24 Bubber P, Haroutunian V, Fisch G, Blass JP, Gibson GE. Mitochondrial abnormalities in Alzheimer brain: mechanistic implications. *Ann Neurol* 2005; **57**: 695–703.
25 Jens B, Maureen B, Stephen F, Karl HW, Robert WM, Yadong H. O3–03–03: Apolipoprotein E4 and its fragment impair mitochondrial dynamics in neuronal cultures. *Alzheimers Dement* 2008; **4**: T163.
26 Mancuso M, Orsucci D, Siciliano G, Murri L. Mitochondria, mitochondrial DNA and Alzheimer's disease. What comes first? *Curr Alzheimer Res* 2008; **5**: 457–468.
27 Roses AD, Saunders AM, Huang Y, Strum J, Weisgraber KH, Mahley RW. Complex disease-associated pharmacogenetics: drug efficacy, drug safety, and confirmation of a pathogenetic hypothesis (Alzheimer's disease). *Pharmacogenomics J* 2007; **7**: 10–28.
28 Chang S, Ma Tr, Miranda RD, Balestra ME, Mahley RW, Huang Y. Lipid- and receptor-binding regions of apolipoprotein E4 fragments act in concert to cause mitochondrial dysfunction and neurotoxicity. *Proc Natl Acad Sci USA* 2005; **102**: 18694–18699.
29 Humphries AD, Streimann IC, Stojanovski D, Johnston AJ, Yano M, Hoogenraad NJ *et al.* Dissection of the mitochondrial import and assembly pathway for human Tom40. *J Biol Chem* 2005; **280**: 11535–11543.
30 Anandatheerthavarada HK, Biswas G, Robin M-A, Avadhani NG. Mitochondrial targeting and a novel transmembrane arrest of Alzheimer's amyloid precursor protein impairs mitochondrial function in neuronal cells. *J Cell Biol* 2003; **161**: 41–54.
31 Devi L, Prabhu BM, Galati DF, Avadhani NG, Anandatheerthavarada HK, Devi L *et al.* Accumulation of amyloid precursor protein in the mitochondrial import channels of human Alzheimer's disease brain is associated with mitochondrial dysfunction. *J Neurosci* 2006; **26**: 9057–9068.
32 Hahn MW, Rockman MV, Soranzo N, Goldstein DB, Wray GA. Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. *Genetics* 2004; **167**: 867–877.
33 Clement M, Posada D, Crandall KA. TCS: a computer program to estimate gene genealogies. *Mol Ecol* 2000; **9**: 1657–1659.
34 Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 2009; **25**: 1451–1452.
35 Tachmazidou I, Verzilli CJ, De Iorio M. Genetic association mapping via evolution-based clustering of haplotypes. *PLoS Genet* 2007; **3**: e111.
36 Cummings CJ, Zoghbi HY. Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum Mol Genet* 2000; **9**: 909–916.
37 Chen S, Ferrone FA, Wetzel R. Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation. *Proc Natl Acad Sci USA* 2002; **99**: 11884–11889.
38 Swami M, Hendricks AE, Gillis T, Massood T, Mysore J, Myers RH *et al.* Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum Mol Genet* 2009; **18**: 3039–3047.
39 Paranjape S, Zeitlin P. Atypical cystic fibrosis and CFTR-related diseases. *Clin Rev Allergy Immunol* 2008; **35**: 116–123.
40 Rave-Harel N, Kerem E, Nissim-Rafinia M, Madjar I, Goshen R, Augarten A *et al.* The molecular basis of partial penetrance of splicing mutations in cystic fibrosis. *Am J Hum Genet* 1997; **60**: 87–94.
41 Bekris LM, Millard SP, Galloway NM, Vuletic S, Albers JJ, Li G *et al.* Multiple SNPs within and surrounding the apolipoprotein E gene

influence cerebrospinal fluid apolipoprotein E protein levels. *J Alzheimers Dis* 2008; **13**: 255–266.
42 Bekris L, Galloway N, Montine T, Schellenberg G, Yu C. APOE mRNA and protein expression in postmortem brain are modulated by an extended haplotype structure. *Am J Med Genet B Neuropsychiatr Genet* 2009. 10.1002/ajmg.b.30993.
43 Park K, Goto K. A review of the benefit-risk profile of gefitinib in Asian patients with advanced non-small-cell lung cancer. *Curr Med Res Opin* 2006; **22**: 561–573.
44 Jiang H. Overview of gefitinib in non-small cell lung cancer: an Asian perspective. *Jpn J Clin Oncol* 2009; **39**: 137–150.
45 Kitada M. Genetic polymorphism of cytochrome P450 enzymes in Asian populations: focus on CYP2D6. *Int J Clin Pharmacol Res* 2003; **23**: 31–35.
46 Faison WE, Schultz SK, Aerssens J, Alvidrez J, Anand R, Farrer LA *et al.* Potential ethnic modifiers in the assessment and treatment of Alzheimer's disease: challenges for the future. *Int Psychogeriatr* 2007; **19**: 539–558.
47 Roses AD. The medical and economic roles of pipeline pharmacogenetics: Alzheimer's disease as a model of efficacy and HLA-B(*)5701 as a model of safety. *Neuropsychopharmacology* 2009; **34**: 6–17.
48 Jiang Q, Heneka M, Landreth GE. The role of peroxisome proliferator-activated receptor-gamma (PPARgamma) in Alzheimer's disease: therapeutic implications. *CNS Drugs* 2008; **22**: 1–14.
49 Risner ME, Saunders AM, Altman JF, Ormandy GC, Craft S, Foley IM *et al.* Efficacy of rosiglitazone in a genetically defined population with mild-to-moderate Alzheimer's disease. *Pharmacogenomics J* 2006; **6**: 246–254.

Supplementary Information accompanies the paper on the The Pharmacogenomics Journal website (http://www.nature.com/tpj)